图书情报工作
第62卷 第13期 2018年7月



# 基于SNA和DMR方法的高血压主题探测与演化趋势比较研究*

■ 周利琴[1]　徐健[1]　巴志超[1]　张斌[2,3]
[1] 武汉大学信息资源研究中心　武汉 430072　[2] 武汉大学中国传统文化研究中心　武汉 430072
[3] 武汉大学国家文化发展研究院　武汉 430072

**摘要**：[**目的/意义**]探测高血压医学文献的主题和演化趋势，对发现高血压领域的研究热点和前沿，理解高血压领域概况和促进专家之间的知识交流具有重要意义。[**方法/过程**]以PubMed数据库下载的26 717篇与高血压相关的文献题录数据作为研究对象，抽取高频主题词构造共现矩阵，同时采用社会网络分析(SNA)和狄利克雷多项回归(DMR)主题模型从中观、微观层面探测高血压医学文献的主题分布和演化趋势；比较这两种方法的关联和异同点。[**结果/结论**]研究发现，高血压医学文献主要集中在危险因素、研究方法、基本要素、诊断治疗和动物实验这5个研究主题，主题的相对分布比率随着时间变化而不断改变。利用SNA方法获取的主题词更加具体和明确，而DMR方法获取的主题词更加宽泛，但在探索各个主题的演化趋势方面比较有优势。

**关键词**：高血压　主题探测　SNA　DMR　主题模型　演化趋势
**分类号**：G250
**DOI**：10.13266/j.issn.0252-3116.2018.13.011



## 1 引言

高血压是最常见的慢性病，也是心脑血管疾病的最重要危险因素。50年来，我国高血压患者逐年递增，目前我国约有2亿高血压患者，每5个成年人中有1人患高血压[1]。高血压的致残、致死率极高，其并发症能导致一半左右的中风和心脏病患者死亡，全球每年有940万人死于高血压[2]。我国高血压患者死亡人数占总死亡人数的46%，心血管疾病占26.9%，严重消耗了医疗资源和社会资源，给家庭、社会和国家都造成了沉重负担。因此，越来越多的学者、互联网巨头、医药巨头们开始投身于高血压领域的研究。而与高血压相关的生物医学文献，作为医学知识传播和继承的载体，其中隐含了大量有用、潜在的信息。但是生物医学文献的数量正在呈爆炸式增长，仅在PubMed数据库中，以“Hypertension”为检索词，就检索到284 322篇与高血压相关的文献，时间截至2017年5月12日。如此大规模的医学文献给科研人员造成了巨大的困扰，使得科研工作者们难以对高血压领域文献的研究进行全貌概览和热点跟进。

很多学者开始尝试使用文献计量学的方法来研究高血压医学文献。文献计量学是一种利用定量和统计分析来描述文献在一个给定的字段或主体中的出版模式[3]。Y. S. Oh和Z. S. Galis[4]利用引文分析和内容分析等文献计量学方法，识别和验证了近百年来发表的关于高血压研究引用率最高的100篇文献的关键特征，包括引文排名、出版年份、出版杂志、文献类型、国家、资金来源和作者身份等。C. Schreiber等[5]利用文献计量学研究了2000-2014年间出版的关于肺动脉高血压临床治疗方面的文献，以期发现肺动脉高血压的研究特征，影响因素和起源国家。M. Götting等[6]从Web of science中检索到了1900-2015年间关于肺动脉高血压的文献，并对其出版的国家分布、时间分布、作者分布、被引情况以及h-指数等进行了分析。

* 本文系国家自然科学基金国际(地区)合作与交流项目“基于慢病知识管理的智慧养老平台研究”(项目编号：71661167007)、国家自然科学基金重点国际(地区)合作研究项目“大数据环境下的知识组织与服务创新”(项目编号：71420107026)和国家自然科学基金青年项目“心智空间视角下科学知识生成与演化机理研究”(项目编号：71704138)研究成果之一。

作者简介：周利琴(ORCID：0000-0001-5105-2669)，博士研究生，E-mail：zhoulq92@163.com；徐健(ORCID：0000-0002-0230-5137)，博士研究生；巴志超(ORCID：0000-0001-5626-5604)，博士研究生；张斌(ORCID：0000-0002-5591-7874)，博士后，讲师。

收稿日期：2017-10-08　修回日期：2018-03-12　本文起止页码：82-91　本文责任编辑：杜杏叶

但是，这些研究都集中在宏观层面(包括主要研究国家、机构、作者等等)，缺乏对中、微观层面的社群分布和主题演化的深入分析。

探测高血压医学文献的主题分布和演化趋势，对发现高血压领域的研究热点和前沿，理解高血压领域概况和促进专家之间的知识交流具有重要意义。现有的比较流行的主题识别主要集中在两个方向：一种是基于社会网络分析(SNA)方法的主题和社群探测，另一种是基于主题模型的主题识别[7]。本文选取了PubMed生物医学文献数据库中2000－2017年间发表的26 717篇与高血压相关的文献题录数据，分别采用社会网络分析和狄利克雷多项回归(Dirichlet-multinomial Regression,DMR)主题模型方法来探测高血压医学文献的主题分布和演化趋势，对高血压生物实体网络和内容进行深入的分析和描述，探讨某一时间下主题的局部变化，以及显著的高血压生物实体之间的相互作用。同时，对这两种方法获得的主题和演化趋势进行比较，分析这两种方法的关联和异同点。

## 2 相关研究回顾

基于共词分析[8-9]和引文分析[10]的热点主题发现方法已经被众多学者研究，其发展成熟，普及度很高。D. R. Swanson[11]早在1987年就采用词语共现的方法，提出了基于生物医学文献的知识发现，引入ABC理论以潜在的知识实体来挖掘、推断隐藏在文献中没有直接联系的生物实体之间的关联，从而解决生物医学文献的信息孤岛问题。但是，单纯的采用词语共现的方法进行分析，需要研究者具有很强的专业医学知识，并且需要大量的人工操作。因此，很多学者在此基础上做出了很多改进，M. D. Gordon和R. K. Lindsay[12-13]在D. R. Swanson实验的基础上采用词频统计、TF-IDF等信息检索方法来发现不可能被简单、标准的文献索引方法发现的，但是主题之间又有潜在联系的，可能有益于科学研究探索的知识。祝清松和冷伏海[14]采用引文分析方法，以高被引论文为研究对象对文献内容进行抽取和主题识别。然而，通过共词分析和引文分析的主题发现方法存在若干问题：共词分析结果独立于文档，我们通常通过浏览共词矩阵的聚类结果来识别主题内容，每一个主题表现为不同词语的聚类，但是，若仅选取一篇文档，我们则无法探知其中所包含的主题分布，这种分析结果与原文献脱离的特征，降低了主题分析的参考性和准确性；而引文分析方法具有滞后性，难以对年代较新的文档集合进行主题分析，不具有时效性。

社会网络分析方法[15]是在人类学、社会学、心理学、统计学等领域的基础上发展起来的一种研究范式和方法[16]，被广泛应用于数据挖掘、知识管理、信息传播、知识网络[17-18]、数据可视化等研究中。社会网络是指社会行动者及其之间关系的集合，主要是对网络中的各种实体之间的关系结构和属性进行测量、分析和预测[19]。在学术文献中，学者们在阐述某一领域热点时，会采用相同或相近的词语来表达，海量文本数据之间的关系可以通过文本的主题词联系起来，形成巨大的词语网络。引入社会网络分析能够更加清晰地、可视化地展现主题词之间的关系网络，为分析主题词的重要程度、主题词在网络中的位置以及与主题词相关联的词语提供有效的支持[20]。M. L. Wallace等[21]利用两个案例研究证明了将社区发现方法用于研究方向的识别是一种非常理想的思路，它能比传统的共被引分析揭示更多的知识领域的结构细节。因此，采用社会网络分析方法挖掘主题词之间的关系网络以及探测主题社群是一个值得深入研究的话题。

主题模型是机器学习领域基于概率统计模型所提出的主题发现方法。主题模型被广泛应用于自然语言处理[22]、信息检索[23]、文本挖掘[24]等领域。在主题模型中，假设文档集中存在K个潜在主题，主题被表达为词项的概率分布，而文档被表达为主题的概率分布，以词袋表示每篇文档。主题模型起源于潜在语义索引(Latent Semantic Indexing,LSI)[25]，其通过奇异值分解(Singular Value Decomposition,SVD)来表达主题空间，并对其进行语义降维。T. Hoffman进一步在其基础上提出概率潜在语义索引(Probabilistic Latent Semantic Indexing,pLSI)[26]，以概率值来区分文档、主题、词项之间相互关联的大小。直至D. M. Blei[27]提出潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)，主题模型才发展到较为成熟的阶段。当前文献中所提到的主流的主题模型，一般即指LDA及其衍生模型。狄利克雷多项回归主题模型(Dirichlet-multinomial Regression,DMR)是D. Mimno和A. McCallum[28]在D. M. Blei提出的LDA模型的基础上扩展和衍生而来的。该模型在文档－主题分布中包含一个对数线性先验概率，可以通过调节观察到的文档特征，例如作者、出版地点、参考文献和出版日期等，获取不同条件下的主题分布。M. Song等[29]利用DMR主题模型探测老年痴呆症的主题分布和演化趋势，获得了很好的效果。相比于共词分析，在得到主题模型的训练结果后，任意抽取一篇

文档,则可以获知文档中的主题概率分布;相比于引文分析,主题模型中的主体内容表现为词项的概率分布,不具有滞后性,主题模型能够较好的反映"词语-主题-文档"之间的联系。因此,主题模型在热点主题发现方面具有很大优势。

综上所述,相比于传统的共词分析和引文分析方法,社会网络分析方法(SNA)和狄利克雷多项回归主题模型(DMR)方法在探测热点主题方面具有重大优势。本文的主要目的是采用这两种方法从中观和微观层面来探测高血压文献的主题分布和演化趋势,比较这两种方法的关联和差异。

## 3 研究设计

本文的研究思路与框架如图 1 所示。主要分为以下五个步骤:①数据收集与处理:从 PubMed 数据库中收集高血压文献相关的标题、摘要、年份、期刊等信息,然后对摘要数据做分词和去停用词等处理;②基础文献计量分析:将处理后的数据导入 BICOMB[30] 书目共现分析系统中,对其年份分布、期刊分布和主题词分布等做基础的文献计量分析;③高血压文献主题社群探测:根据步骤②中获取的 MeSh 主题词,对其构造主题词共现矩阵,然后将该矩阵导入 Gephi[31] 中,运用社会网络分析方法计算主题词的 pagerank 值和中心度,识别关键节点,然后对主题社群进行探测和可视化展示,并描绘其演化趋势;④采用狄利克雷多项回归主题模型,对其主题分布进行探测,并研究主题随时间演化的趋势;⑤对步骤③、④中得出的分析结果进行比较和验证。

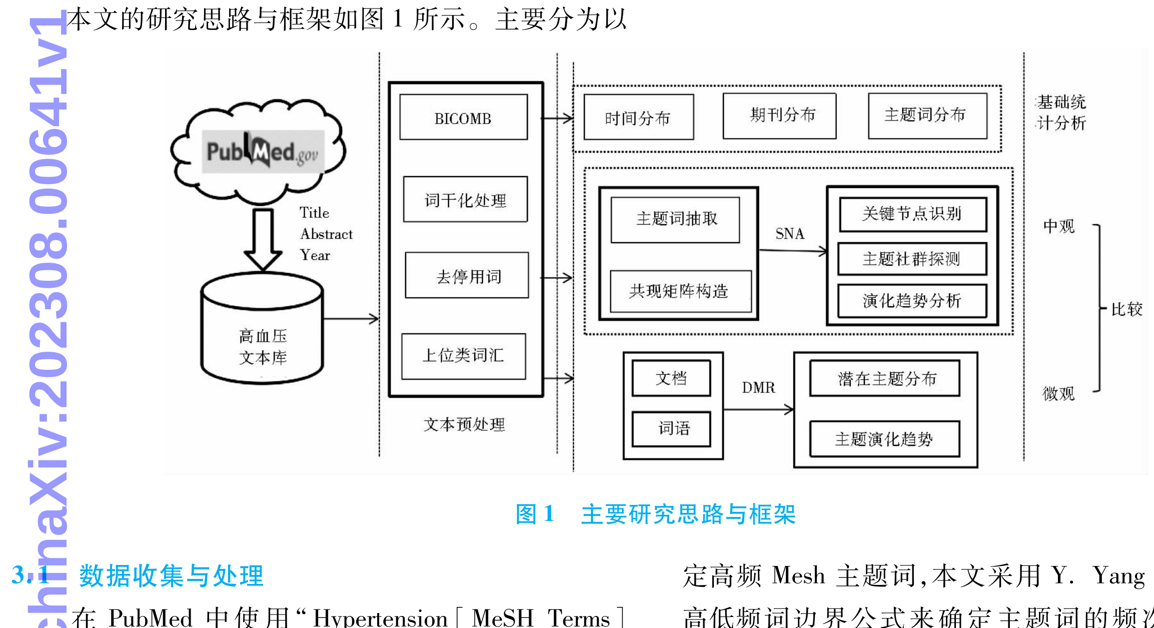



图 1 主要研究思路与框架

### 3.1 数据收集与处理

在 PubMed 中使用"Hypertension[MeSH Terms] AND ("2000/1/1"[PDat]:"2017/5/1")"为检索策略,共检索到 2000 年以来与高血压有关的文献 99 252 篇,选取同时包含摘要和全文的文献题录信息,共计 26 717 篇,检索时间截至 2017 年 5 月,保存为 XML 格式,这是本文的研究对象。

将这些数据导入书目共现分析系统(Bibliographic Items Co-occurrence Matrix Builder,BICOMB)中,该系统可对 PubMed 数据库、引文数据库 SCI、中国知网 CNKI 等数据库中的书目信息进行快速读取、准确提取字段并归类存储、统计,并生成书目数据的共现矩阵。通过对期刊、作者、主题词等字段的提取和统计,发现本文获取的 26 717 篇文献分布在 1 701 种期刊,涉及 171 637个作者,9 978 个主题词。

### 3.2 方法与模型

(1)高频主题词抽取和共现矩阵的构造。为了确定高频 Mesh 主题词,本文采用 Y. Yang 等[32]使用的高低频词边界公式来确定主题词的频次阈值,如下所示:

$$T=(-1+\sqrt{1+8*I_1}/2) \quad (1)$$

其中,$I_1$ 是仅出现一次的主题词数量,T 是高频词中最小的频次阈值。根据此边界公式获取主题词的频次阈值,从题录数据的摘要中获取高频 Mesh 主题词,并在 BICOMB 系统中构建主题词共现矩阵。

(2)基于优化网络模块度的社群探测。社群是由一群高度聚集、联系紧密的节点聚集组成的,是一种介于宏观和微观之间的网络特征,也是社会网络中的常见现象[33]。对于真实网络,属于同一个社群的节点更有可能具有相似或相近的功能,社群结构可以帮助人们理解网络结构和功能之间的关系。最具代表性的社群识别算法是 M. E. J. Newman[34]提出的基于优化网络模块度(Modularity)方法,模块度是一种可以衡量网络划分好坏的指标,也叫 Q 值,其计算方法如下:

$$Q = \sum_i (e_{ij} - a_i^2) \qquad (2)$$

其中，$e_{ij}$表示社群 $i$ 和社区 $j$ 之间的边数占总边数的比率；$a_i = \sum_j e_{ij}$表示有一个端点在社群 $i$ 中的边占边总数的比率。从本质上来说，基于模块度的算法是根据边的中介性和模块度的变化进行社区识别。由于共词网络中的节点就是主题词，确定社群代表主题的过程就转化为寻找核心节点的过程，少数核心节点代表了社群对应的科研主题。在复杂网络中，节点的重要性指标有很多，除了传统的中心度、声望等指标，还有 PageRank 值。这些指标都从网络全局层面，考虑计算每一个节点在整个网络中的边数、中心性以及与其他节点的连接情况，从而判断出核心节点。

（3）狄利克雷多项回归主题模型。狄利克雷多项回归主题模型是 D. Mimno 和 A. McCallum[28] 在 D. M. Blei 提出的 LDA 模型的基础上扩展和衍生而来的。该模型主要通过调节观察到的文档特征来获取不同条件下的主题分布，本文把高血压文献出版的时间作为变量，探讨主题随着时间的变化趋势。DMR 主题模型图如图 2[28] 所示：

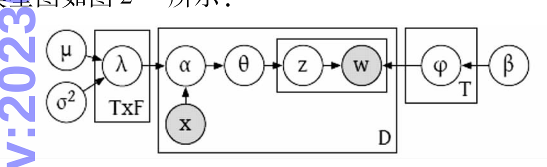



图 2　狄利克雷多项回归主题模型

在文档集合 D 中，对于每一个文档 $d$，$x_d$ 表示元数据的特征向量，α 是可观察的文档特征的函数，表示主题的先验概率分布；给定先验概率分布 N(0,Σ)，超参数 β，文档和词语的生成过程如下：

For each topic $t$，draw $\varphi_t \sim Dir(\beta)$. $Dir(\beta)$是与先前的狄利克雷分布不同的主题－词语分布；

For each document $d$，draw $\theta_d \sim Dir(\alpha_d) = \mathrm{Dir}(\exp(\tau_d))$，$\tau_d \in \tau$. 对于每一个文档的 $\alpha_d$，狄利克雷分布的参数和 $\tau_d$ 是协方差函数 $f(y_d, x_k)$，其中 $y_d$ 是文档 d 的观察属性向量，$x_k$ 是元数据的向量；

For each word $w$，draw $z_{d,w} \sim Multi(\theta_d)$. $z_{d,w}$是词语 $t_w$ 的主题分配，$\theta_d$ 是文档 $d$ 属于某个主题的比例；draw $T_{d,w} \sim Multi(\varphi_{z_{d,w}})$. $T_{d,w}$是文档 $d$ 的第 $w$-$th$ 个单词，$\varphi_t$ 是主题 $t$ 的偏好，$\Sigma_n \varphi_{t,n} = 1$。

在 DMR 主题模型中，我们设置三个固定参数：$\sigma^2$，先前分布的参数值的方差；$\beta$，狄利克雷主题－词语分布；|T|，主题的个数。

## 4　分析与结果

### 4.1　基于 SNA 方法的高血压热点主题探测与演化趋势分析

根据高低频词边界公式（1）计算，可得到频次边界为 77。为了更好的实现可视化，我们删除了频次阈值为 77 以下的点和与其他节点没有链接的节点，共得到 632 个顶点，对高频词构造共现矩阵。为减少其复杂度，选取前 100 个顶点，可得到 4 950 条边。顶点是指由文章衍生而来的生物实体，边表示实体之间的关系，边的权重表示两个实体在文章特定句子中共同出现的频率。导入 Gephi，利用社区探测算法[35] 对其进行可视化，如图 3 所示：

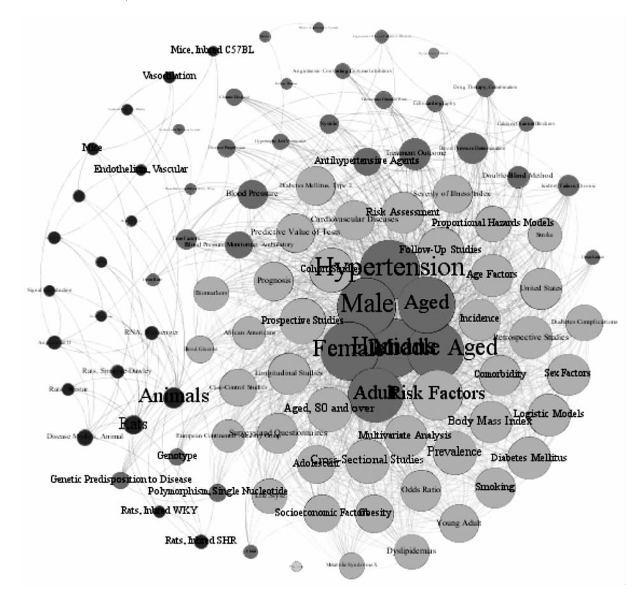



图 3　高血压文献社群和热点主题探测

（1）关键节点识别。为了识别高血压医学文献中最核心的生物实体，我们以 4 个著名的中心指标来分析前 10 个中心节点：pagerank 值、加权中心性、接近中心性、中间中心性，见表 1。详细的中心度测量是通过 S. Wasserman [15] 和 S. Brin [36] 提出的方法。

PageRank 是基于其传入连接数量的总和来估计一个特定节点的重要性[36]。表 1 展示了 PageRank 值排名前 10 的实体。平均 PageRank 值是 0.01，PageRank 值排名靠前的生物实体类似于加权中心度排名靠前的实体，10 个实体有 8 个一样，但是排名顺序有变化。而且，PageRank 有独一无二的生物实体，比如说 rates 仅出现在排名前 10 的 PageRank 值中。

特定节点的度中心度是指连接到该节点的边的数量，而加权中心度是度中心度的扩展，通过某一个特定节点的每个节点对的频率来计算；中间中心度被定义

为通过给定节点的最短路径的数量;接近中心度表示从一个特定节点到网络中所有其他节点的总距离之和的倒数,表示该节点在网络中的可扩展性[15]。表 1 描述了各中心度排名前 10 的生物实体。平均加权度是 23.427;平均中间中心度是 78.17,平均接近中心度是 0.4。加权中心度和中间中心度排名前 10 的节点中,有 7 个相同,risk factor、Prospective Studies 和 Follow-up studies 这三个节点仅出现在加权中心度中,而 Blood Pressure、Animals 和 Body Weight 仅出现在中间中心度中;接近中心度排名前 10 的词与加权中心度完全相同。

总的来说,在整个网络中处于关键位置的节点主要有 Hypertension、Male、Female、Age、Adult、Human、Risk factors、Body Weight、Blood Pressure、Animals、Prospective Studies、Follow-up studies……等,这些节点在社群分布图中也都处于比较核心的位置。

表 1 PageRank、加权中心性、接近中心性和中间中心性排名 Top-10 节点

| Bio-entity | Page ranks | Bio-entity | weighted degree | Bio-entity | betweeness centrality | Bio-entity | closness centrality |
|---|---|---|---|---|---|---|---|
| Animals | 0.028 2 | Hypertension | 62.171 0 | Blood Pressure | 1 794.118 4 | Hypertension | 0.595 1 |
| Rats | 0.023 2 | Humans | 60.1870 | Animals | 1489.7500 | Male | 0.584 3 |
| Hypertension | 0.022 2 | Female | 60.125 0 | Body Weight | 1 482.000 0 | Humans | 0.577 4 |
| Male | 0.020 3 | Male | 60.0320 | Hypertension | 449.105 9 | Female | 0.577 4 |
| Humans | 0.019 8 | Middle Aged | 59.026 0 | Male | 273.786 6 | Middle Aged | 0.574 0 |
| Female | 0.018 9 | Aged | 58.248 0 | Humans | 211.524 5 | Aged | 0.574 0 |
| Middle Aged | 0.018 5 | Adult | 56.407 0 | Female | 211.524 5 | Adult | 0.567 3 |
| Aged | 0.018 5 | Risk Factors | 51.198 0 | Middle Aged | 189.574 5 | Risk Factors | 0.548 0 |
| Adult | 0.017 8 | Prospective Studies | 45.020 0 | Aged | 189.574 5 | Prospective Studies | 0.533 0 |
| Risk Factors | 0.016 1 | Follow-Up Studies | 44.983 0 | Adult | 166.225 9 | Follow-Up Studies | 0.533 0 |

(2)主题社群探测。根据第 3 节中提出的最优模块度公式(2),使用 V. D. Blondel[35] 提出的模块化算法进行社群探测,并将分辨率 resolution 设为 1[37]。可以得到 5 个模块,其中 3 个主要模块(如图 3 中红色、蓝色和绿色所示),模块化值 Q 是 0.187;平均加权度为 23.43;图密度为 0.282,可见度是 28.2%;平均聚类系数是 0.808,迭代次数 100;特征向量中心度 0.002 39;平均路径长度 2.645;直径是 6。选择 Fruchterman Reingold 布局,图中浅蓝色和紫色的两个社区仅占了总社区的 1%,可忽略不计。因此,我们主要考虑图中红色、蓝色和绿色的三个社区。最大的绿色社区占了整个网络的 42%,包含 Risk Factors、Aged, 80 and over、Prevalence 、Sex Factors 、Prospective Studies、Follow-Up Studies、Aged, 80 and over、Cross-Sectional Studies、Cohort Studies……等词;第二大社区(红色部分)占整个网络的 36%,包含 Hypertension 、humans、female、middle aged、Treatment Outcome、Antihypertensive Agents、Blood Pressure Determination……等词;第三大社区(蓝色部分)占整个网络的 20%,包含 animals、Rats, Inbred SHR、RNA, Messenger、Mice……等词。通过人工判断和专家识别,初步可将这三个社区分成五个主题。

主题 1 主要是包含与高血压危险因素相关的词,比如年龄、怀孕、性别、抽烟、肥胖、心梗等等;主题 2 主要包含高血压相关的研究方法和模型,比如前瞻性研究、随访研究、横向研究、队列研究、回顾性研究等等;还包括高血压研究的一些指标参数,例如发病率、疾病严重程度指数、实验预期值等等。主题 3 主要包括高血压的基本要素,比如性别、年龄、血压、心率、肾素,肾小球滤过率等等;主题 4 主要包括疾病诊断的结果,比如诊断效果、降压药、血压测定、药物剂量反应关系等等。主题 5 主要包含动物、大鼠、RNA、Inbred SHR 等指标,即通过动物实验对高血压的各项指标进行验证。总的来说,每个社区内的生物实体都紧密联系在一起,形成各种不同的与高血压文献相关的特定研究主题。高血压文献主题和社群分布见表 2。

(3)演化趋势分析。将高血压文献按照时间分布分为三个阶段:2000-2005 年、2006-2010 年、2011-2017 年,经过处理后将其分别导入 Gephi,利用社区探测算法对其进行可视化,探测到的各阶段的主题社群分布如图 4(a)、4(b)和 4(c)所示,具体的主题分析方法同上,由于篇幅有限,在此不做详述。

表 2　基于 SNA 的高血压文献社群和主题分布

| community 1 | | community 2 | | community 3 |
|---|---|---|---|---|
| topic 1 | topic 2 | topic 3 | topic 4 | topic 5 |
| 危险因素 | 研究方法 | 基本要素 | 诊断治疗 | 动物实验 |
| Risk Factors 危险因素 | Prospective Studies 前瞻研究 | Hypertension 高血压 | Treatment Outcome 治疗结果 | Animals 动物 |
| Age Factors 年龄因素 | Follow-Up Studies 后续研究 | Humans 人 | Antihypertensive Agents 抗高血压药物 | Rats 大鼠 |
| Sex Factors 性别因素 | Cross-Sectional Studies 交叉研究 | Female 女性 | Blood Pressure Determination 血压测定 | Disease Models, Animal 动物疾病模型 |
| Smoking 吸烟 | Cohort Studies 队列研究 | Male 男性 | Double-Blind Method 双盲方法 | Rats, Inbred SHR 老鼠,自交 |
| Cardiovascular Diseases 心血管疾病 | Retrospective Studies 回顾性研究 | Aged 年龄 | Renin-Angiotensin System 肾素－血管紧张素系统 | Angiotensin II 血管紧张素 |
| Obesity 肥胖症 | Multivariate Analysis 多元分析 | Blood Pressure 血压 | Dose-Response Relationship, Drug 剂量反应关系 | RNA, Messenger RNA 信使 |
| Diabetes Complications 糖尿病并发症 | Logistic Models 逻辑模型 | Heart Rate 心率 | Drug Therapy, Combination 药物治疗组合 | Rats, Sprague-Dawley |
| Life Style 生活方式 | Longitudinal Studies 纵向研究 | Renin 肾素 | Kidney Failure, Chronic 慢性肾衰竭 | Mice 老鼠 |
| Stroke 中风 | Surveys and Questionnaires 调查问卷 | Sympathetic Nervous System 交感神经系统 | Time Factors 时间因素 | Rats, Wistar |
| Body Mass Index 体重指数 | Case-Control Studies 个案控制研究 | Genotype 基因型 | Disease Progression 疾病进展 | Nitric Oxide 一氧化氮 |
| | Proportional Hazards Models 比例危险模型 | Systole 收缩压 | Polymorphism, Single Nucleotide 多态性,单核苷酸 | Rats, Inbred WKY |
| | Incidence 发生率 | Glomerular Filtration Rate 肾小球滤过率 | Genetic Predisposition to Disease 疾病遗传易感性 | Signal Transduction 信号转导 |
| | Risk Assessment 风险评估 | Echocardiography 超声心电图 | Hypertrophy, Left Ventricular 左心室肥大 | Oxidative Stress 氧化应激 |
| | Severity of Illness Index 严重疾病指数 | Calcium Channel Blockers 钙通道阻滞剂 | Heart Failure 心脏衰竭 | Sodium Chloride, Dietary 氯化纳、膳食 |
| | Predictive Value of Tests 试验预测值 | Body Weight 体重 | | Mice, Inbred C57BL |

根据可视化的分布图以及各项参数可知,三个阶段的高血压主题都分布在三个社群,各社群占比相对比较平均,2000－2005 年间各社群占比为 38%、33%、29%,2006－2010 年间各社群占比为 40%、38%、22%,2011－2017 年间各社群占比为 42%、37%、21%;对比三个阶段获得的 MeSH 主题词,发现:三个阶段中出现频次最高的都是 hypertension、male、female 等词,大部分主题词同时出现在三个阶段中,但是各阶段主题社群分布略有差异。各阶段的主题社群分布参数如表 3 所示。这三个阶段中的节点平均度、图密度和模块化参数都在不断增大,说明随着时间的增长,各阶段的 MeSH 主题词数量不断增多,主题社群分布也不断发生变化。但是这种方法工作量很大,并且难以准确探测各个主题在每一个时间段的占比和主题随时间演化的路径。

表 3　基于 SNA 方法的主题社群分布参数

| | 2000－2005 年 | 2006－2010 年 | 2011－2017 年 |
|---|---|---|---|
| 平均度 | 13.94 | 16.88 | 20.4 |
| 图密度 | 0.141 | 0.171 | 0.202 |
| 模块化 | 0.126 | 0.158 | 0.175 |
| 平均聚类系数 | 0.891 | 0.438 | 0.869 |

### 4.2　基于 DMR 方法的高血压热点主题探测与演化趋势分析

(1)热点主题分布。将 26 717 篇与高血压相关的文献题录数据做如下处理:①词干化处理 stemming;②去停用词和长度为 1 的词,出现频率少于 5 次的词语;③去高血压领域的上位类词汇;每篇题录数据生成一个文本文件,作为 DMR 主题模型的文档。然后根据第 3 节中介绍的 DMR 模型和算法,通过一个开源的机器学习语言处理包 Mallet[38] 对数据进行处理。为了与 4.1 节中探测到的主题形成对比,将主题个数|T|设为

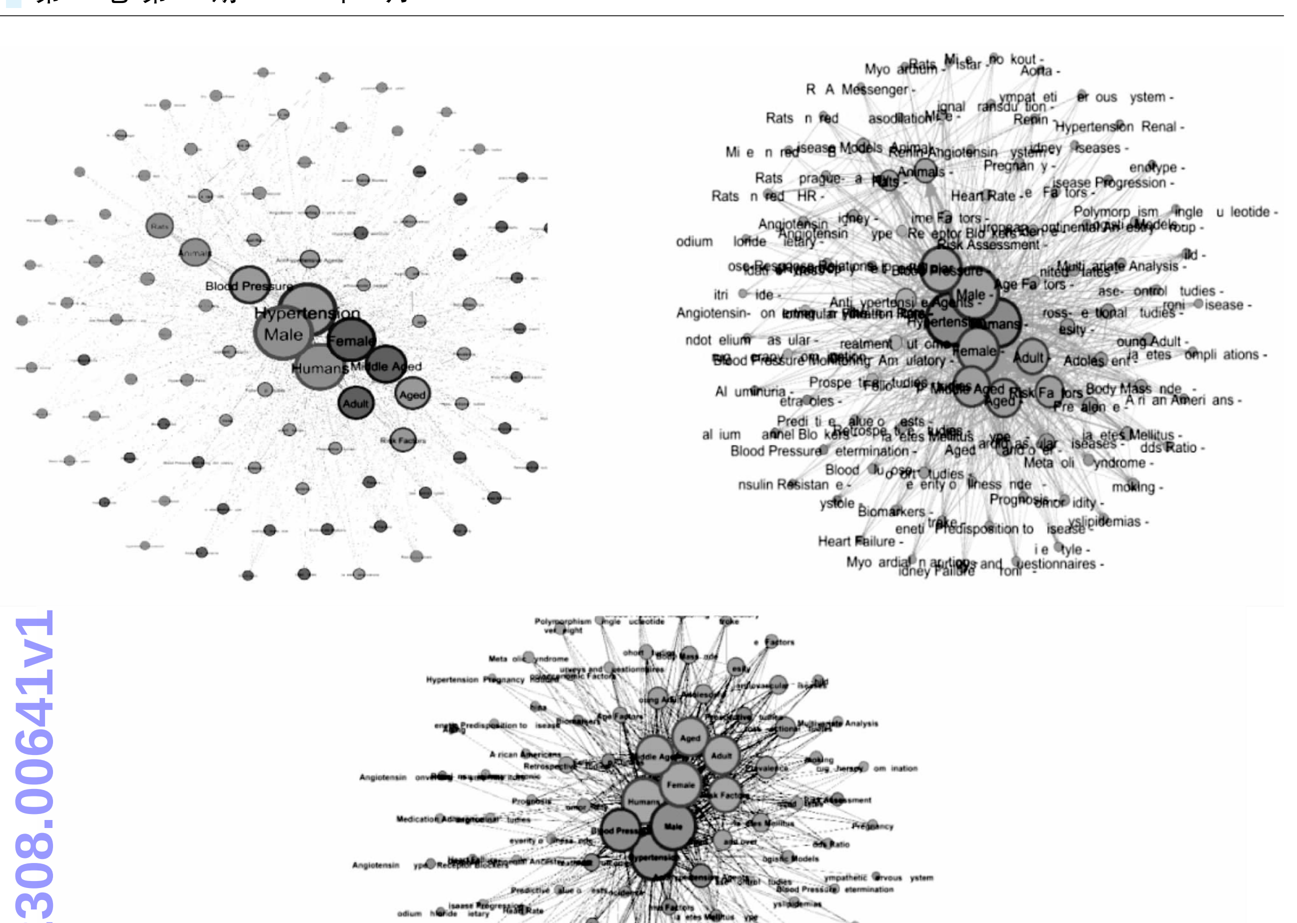





图 4　各阶段主题社群分布

（a）2000 –2005 年主题社群分布；（b）2006 –2010 年主题社群分布；（c）2011 –2017 年主题社群分布

5，同时调节 $\sigma^2$ 和 $\beta$ 值，最终形成 5 个主题和相关主题词。为了使识别到的主题和主题词更有意义，分别将每个主题展示的主题 – 词汇分布中的主题词设为 10 个、20 个、30 个，选取 10 个出现比较频繁并且有意义的词，见表 4。

主题 1 包含 mice、angiotensin、renin、vacular、response、effects、receptor、rats 等词，主要是要来描述与高血压相关的动物实验；主题 2 包含 risk、factors、age、obesity、gene、women 等词，主要用来描述高血压的危险因素，包括年龄、糖尿病、肥胖、性别、基因等等；主题 3 包含 systolic、diatolic、group、compare、rate、invalid、significant 等词，主要用来描述与高血压相关的研究方法；主题 4 包含 patient、blood、gene、treatment、results 等词，主要用来描述高血压基本要素；主题 5 包含 treatment、antihypertensive、theropy、coronary、mortality、medication、care、control 等词，主要用来描述高血压的诊断和治疗。

（2）主题演化趋势。然后，将高血压文献出版的时间作为变量，探讨主题随着时间的变化趋势。统计 2000 年到 2017 年间，每个主题的相对分布情况，如下图 5 所示。总的来说，随着时间的推移，每个主题都在不断发展变化。在 2000 年的时候，主题 1（动物实验）和主题 4（基本要素）所占的比重比较大，主题 5（诊断治疗）的研究相对比较薄弱；随着时间的推移，主题 1（动物实验）呈逐渐下降的趋势，主题 4（基本要素）呈先下降再上升的趋势，但是在 5 个主题中，一直处于比较重要的位置；而主题 5（诊断治疗）的比重逐年增加，到 2017 年已经占有相对重要的比重；主题 2（危险因素）发展比较平稳，一直处于相对重要的位子；主题 3（研究方法）稍有波动，从 2007 年开始比重逐年增加。

**表 4　基于 DMR 的高血压文献主题分布**

| Topic 1<br>动物实验 | Topic 2<br>危险因素 | Topic 3<br>研究方法 | Topic 4<br>基本要素 | Topic 5<br>诊断治疗 |
|---|---|---|---|---|
| Mice | Risk | Systolic | Patient | Patient |
| Renal | Age | Group | Blood | Antihypertensive |
| Vascular | Prevalence | Significant | Treatment | Coronary |
| Effects | Blood | Rate | Results | medication |
| Expression | Disease | Pressure | Finding | clinical |
| Angiotensin | Factors | Diatolic | cardiovascular | Treatment |
| Proteins | Diabetes | Compare | Gene | Theropy |
| Response | Obesity | Invalid | Association | Mortality |
| Receptor | Gene | Term | Evidence | care |

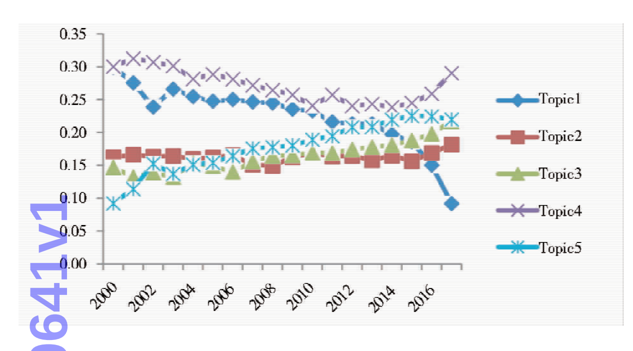



图 5　基于 DMR 方法的高血压文献主题演化趋势

## 5　讨论:SNA 和 DMR 方法比较分析

结果发现,基于 SNA 方法获取的主题与基于 DMR 主题模型探测到的主题基本相同,都包括危险因素、研究方法、基本要素、诊断治疗和动物实验这 5 个研究主题。从宏观上看,SNA 方法获得的主题中,主题词 Mesh Terms 更加具体化,每个社群或主题内的主题词意思比较明确,而 DMR 方法获得的主题中,主题词相对比较宽泛。比如说在主题危险因素中,SNA 方法识别出的主题词包括 age factors、Diabetes Mellitus、sex factors、smoking、cardiovascular disease、obesity、life style 等等,这些词都能代表比较具体的危险因素;而 DMR 方法识别出的危险因素的主题词主要包括 age、Diabetes、obesity、gene 等,这些词更宽泛,仅仅只能够代表危险因素的各个大类。另外,在主题研究方法中,SNA 识别出的主题词不仅包括 prospective studies、logistic models、surveys and questionnaires、cross-sectional studies 等表示研究方法的词,还包括 risk assessment、odds ratio、severity of illness index 等指标;而 DMR 方法识别出的主题词主要包括 group、compare、rate、significant 等比较宽泛的词。产生这种结果的原因,一方面是由于数据集太大,SNA 方法选取的研究对象是出现频率最高的 TOP－100 个主题词,而 DMR 方法采用的是整个文档集合,研究对象不同,产生的结果也会产生差异;另一方面是 SNA 方法获取的社群主题数目是人为主观确定的,而 DMR 方法也需要预先设定主题和主题词的个数,这样会导致一定的误差。另外,DMR 方法可以较好的反映“词语－主题－文档”之间的联系,任意抽取一篇文档都可以获知文档中的主题概率分布。

在演化趋势的比较方面,基于 SNA 方法只能探测每一个时间段的主题社群分布,难以比较各个阶段主题发生的变化和演化趋势;而基于 DMR 主题模型的方法可以探测不同主题在每一个时间段的占比,以及主题随时间的演化情况,在探索主题演化趋势的过程中也比较有优势。为了将通过 SNA 方法和 DMR 方法获得的主题和演化进行对比分析,本文在 4.1 节和 4.2 节中均将探测的主题数设为 5 个。但是,这种人为设置主题数目和主观判断的方式可能存在一定的误差,这也是我们下一步需要解决和探讨的问题。

总的来说,SNA 方法和 DMR 主题模型在探测慢病社群和主题演化趋势方面有不同侧重,SNA 方法获取的主题词更加具体和明确,而 DMR 方法获取的主题词比较宽泛,需要人工解读,但是在探索各个主题的演化趋势方面比较有优势。若将二者结合起来,同时从中、微观层面探索知识网络的社群和主题演化趋势,即可相辅相成。

## 6　结语

本文分别从中观层面采用社会网络分析方法探测了高血压文献的主题社群分布和演化趋势,从微观层面采用狄利克雷多项回归主题模型探索了高血压文献的主题分布及其演化趋势,最后还比较了 SNA 和 DMR 两种方法的关联与优缺点。

研究发现:①高血压领域的研究文献总的来说分为危险因素、研究方法、基本要素、诊断治疗和动物实验这 5 个研究主题;②随着时间推移,每个主题都在不断变化,基本要素的研究所占比重一直比较大,动物实验的研究逐年减少,危险因素的研究发展比较平稳,一直处于相对重要的比重,而研究方法的研究稍有波动,从 2007 年开始比重逐年增加;③SNA 和 DMR 方法探测的

主题基本相同,但主题词略有差异,从宏观上看,SNA 主题识别效果更好,主题词更加集中,但是 DMR 在探测主题演化趋势方面有优势,二者可结合使用,效果更佳。

这些研究可以帮助刚刚接触高血压领域的研究者了解该领域概况,发现该领域的研究热点和预测研究前沿,促进领域专家之间进行领域内部和跨领域的知识交流,帮助决策者跟进高血压领域知识的流动情况。同时,本文中的社区探测和主题演化趋势的分析方法可以扩展到慢病其他领域,比如说糖尿病、心血管疾病、冠心病等等。

本文有一定的局限性,DMR 主题模型在探测潜在主题的时候,需要预先设定主题的数量,通过文档困惑度[39]确定的主题数目比较大,在此不合适,因此,本文 DMR 识别的主题数量是主观确定的,可能存在一定的误差;另外,本文还缺少对社群或者主题内部结构的关联探测,这也是下一步需要做的工作。



**参考文献:**

[1] 中国高血压防治指南修订委员会. 中国高血压防治指南 2010[J]. 中华心血管病杂志, 2011, 39(7):701 –708.

[2] NATIONAL INSTITUTES OF HEALTH. Report: estimates of funding for various research, condition, and disease categories (RCDC)[EB/OL]. [2017 –06 –02]. http://report.nih.gov/categorical_spending.aspx.

[3] SONG M, KIM S, ZHANG G, DING Y, et al. Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. Journal of the association for information science and technology, 2014, 65(2), 352 –371.

[4] OH Y S, GALIS Z S. Anatomy of success: the top 100 cited scientific reports focused on hypertension research [J]. Hypertension, 2014, 63(4):641 –647.

[5] SCHREIBER C, EDLINGER C, EDER S, et al. Global research trends in the medical therapy of pulmonary arterial hypertension in 2000 –2014 [J]. Pulmonary pharmacology & therapeutics, 2016, 39(8):21 –27.

[6] GOTTING M, SCHWARZER M, GERBER A, et al. Pulmonary hypertension: scientometric analysis and density-equalizing mapping [J]. Plos one, 2017, 12(1): e0169238.

[7] DING Y. Community detection: Topological vs Topical [J]. Journal of informetrics, 2011, 5(4) : 498 –514.

[8] KLAVANS R, BOYACK K W. Identifying a better measure of relatedness for mapping science[J]. Journal of the American Society for Information Science and Technology, 2006, 57(2): 251 –263.

[9] RONDA-PUPO G A, GUERRAS-MARTIN L A. Dynamics of the evolution of the strategy concept 1962 –2008: a co-word analysis [J]. Strategic management journal, 2012, 33(2):162 –188.

[10] CHEN C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American society for Information Science and Technology, 2006, 57(3): 359 –377.

[11] SWANSON D R. Two medical literatures that are logically but not bibliographically connected [J]. Journal of the Association for Information Science. 1987;228 –233.

[12] LINDSAY R K, GORDON M D. Literature-based discovery by lexical statistics[J]. Journal of the American Society for Information Science and Technology, 1999, 50(7):574 –587.

[13] GORDON M D, LINDSAY R K. Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil [J]. Journal of the Association for Information Science and Technology, 1996, 47(2):116 –128.

[14] 祝清松, 冷伏海. 基于引文内容分析的高被引论文主题识别研究[J]. 中国图书馆学报, 2014, 40(1):39 –49.

[15] WASSERMAN S, FAUST K. Social network analysis: methods and applications. [J]. Contemporary sociology, 1994, 91(435):219 –220.

[16] 朱庆华, 李亮. 社会网络分析法及其在情报学中的应用[J]. 情报理论与实践, 2008, 31(2):179 –183.

[17] 徐媛媛, 朱庆华. 社会网络分析法在引文分析中的实证研究[J]. 情报理论与实践, 2008, 31(2):184 –188.

[18] 李亮, 朱庆华. 社会网络分析方法在合著分析中的实证研究[J]. 情报科学, 2008, 26(4):549 –555.

[19] BUTTS C T. Social network analysis: a methodological introduction [J]. Asian journal of social psychology, 2008, 11(1):13 –41.

[20] 王洪伟, 高松, 陆颋. 基于 LDA 和 SNA 的在线新闻热点识别研究[J]. 情报学报, 2016, 35(10):1022 –1037.

[21] WALLACE M L, GINGRAS Y, DUHON R. A new approach for detecting scientific specialties from raw cocitation networks [J]. Journal of the American Society for Information Science and Technology, 2009, 60 (2) : 240 –246.

[22] GROSS A, MURTHY D. Modeling virtual organizations with Latent Dirichlet Allocation: a case for natual language processing[J]. Neural networks, 2014,58:38 –49.

[23] TANG X B, Fang X K. Research on the subject retrieval of the weibo based on the integration of text clustering and LDA[J]. Information studies: theory & application, 2013,8: 85 –90.

[24] ZHANG P J, SONG L. Overview on topic modeling method of microblogs text based on LDA[J]. Library and information service, 2012,24:120 –126.

[25] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 2010, 41(6): 391 –401.

[26] HOFMANN T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 1999: 50 –57.

[27] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine Learning research archive, 2003, 3: 993 -1022.

[28] MIMNO D, MCCALLUM A. Topic models conditioned on arbitrary features with Dirichlet-multinomial Regression [J]. University of Massachusetts - Amherst, 2012, 2008: 411 -418.

[29] SONG M, HEO G E, LEE D. Identifying the landscape of Alzheimer's disease research with network and content analysis[J]. Scientometrics, 2015, 102(1):905 -927.

[30] CUI L. Development of a text mining system based on the co-occurrence of bibliographic items in literature databases[J]. New technology of library and information service,2008, 24(8):70 -75.

[31] BASTIAN M, HEYMANN S, JACOMY M. Gephi: an open source software for exploring and manipulating networks[C]//Proceedings of the third international conference on Weblogs and social media. California: ICWSM, 2009.

[32] YANG Y, WU M, CUI L. Integration of three visualization methods based on co-word analysis[J]. Scientometrics, 2012, 90(2): 659 -673.

[33] 程齐凯, 王晓光. 一种基于共词网络社区的科研主题演化分析框架[J]. 图书情报工作, 2013, 57(8):91 -96.

[34] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J]. Physical review E, 2004, 69(2): 026113.

[35] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008 (10):155 -168.

[36] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer networks, 1998, 30(1 -7):107 -117.

[37] LAMBIOTTE R, DELVENNE J C, BARAHONA M. Laplacian dynamics and multiscale modular structure in networks[J]. Physics, 2008,1(2):1 -29.

[38] WALLACH H M, MIMNO D M, MCCALLUM A. Rethinking LDA: Why priors matter[J]. Advances in neural information processing systems, 2009, 23:1973 -1981.

[39] WANG Y, AGICHTEIN E, BENZI M. TM-LDA: efficient online modeling of latent topic transitions in social media[C]//Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. New York:ACM, 2012:123 -131.

**作者贡献说明:**

周利琴:资料收集与整理,论文撰写;

徐健: 数据处理;

巴志超: 研究框架的设计与指导;

张斌: 提供论文修改建议。

## Comparative Analysis of the Topic and Evolution Trend of Hypertension Study Based on SNA and DMR

Zhou Liqin[1] Xu Jian[1] Ba Zhichao[1] Zhang Bin[2,3]

[1] Center for Studies of Information Resources, Wuhan University, Wuhan 430072

[2] Center of Traditional Chinese Cultural Studies, Wuhan University, Wuhan 430072

[3] National Institute of Cultural Development, Wuhan University, Wuhan 430072

**Abstract**: [**Purpose/significance**] Exploring the topic and evolution trend of hypertension literature is of great significance for users to understand the profile, research hot-spots and frontiers of chronic disease, and can promote the knowledge communication among experts. [**Method/process**] This paper takes the Hypertension and 26717 articles from PubMed database as the research object, extracts high-frequency Mesh Terms to construct a co-occurrence matrix. Social network analysis is applied to detect the community and topic distribution of the hypertension study literature, and the expanded topic modeling Dirichlet-multinomial regression is also used to explore the topic distribution and evolution trends. Then similarities and differences of the SNA and DMR method in topic detection are analyzed. [**Result/conclusion**] It is found that the hypertension literature is mainly concentrated on three communities, which can be divided into five research topics, such as risk factors, research methods, basic situation of patients, diagnosis and treatment, and animal experiments. The relative distribution of the topic varies with time change. It is also found that the topic obtained from SNA and DMR are basically similar. But the Mesh Terms obtained from SNA method are more specific and clearer, while the DMR is more broadly and have an advantage in exploring the evolution of various themes.

**Keywords**: hypertension community detection SNA DMR topic model evolution trend